Bilingual lexical extraction based on word alignment for improving corpus search

Jelena Andonovski, Branislava Šandrih, Olivera Kitanović

Bilingual lexical extraction based on word alignment for improving corpus search | Jelena Andonovski, Branislava Šandrih, Olivera Kitanović | The Electronic Library | 2019 | |

10.1108/EL-03-2019-0056

http://dr.rgf.bg.ac.rs/s/repo/item/0005009

Bilingual lexical extraction based on word alignment for improving corpus search

Jelena Andonovski
University of Belgrade, University Library "Svetozar Markovic", Belgrade, Serbia

Branislava Šandrih
University of Belgrade, Faculty of Philology, Belgrade, Serbia, and

Olivera Kitanović
University of Belgrade, Faculty of Mining and Geology, Belgrade, Serbia

Received 1 March 2019
Revised 28 May 2019
13 July 2019
Accepted 30 July 2019

Abstract

Purpose – This paper aims to describe the structure of an aligned Serbian-German literary corpus (SrpNemKor) contained in a digital library Bibliša. The goal of the research was to create a benchmark Serbian-German annotated corpus searchable with various query expansions.

Design/methodology/approach – The presented research is particularly focused on the enhancement of bilingual search queries in a full-text search of aligned SrpNemKor collection. The enhancement is based on using existing lexical resources such as Serbian morphological electronic dictionaries and the bilingual lexical database Termi.

Findings – For the purpose of this research, the lexical database Termi is enriched with a bilingual list of German-Serbian translated pairs of lexical units. The list of correct translation pairs was extracted from SrpNemKor, evaluated and integrated into Termi. Also, Serbian morphological e-dictionaries are updated with new entries extracted from the Serbian part of the corpus.

Originality/value – A bilingual search of SrpNemKor in Bibliša is available within the user-friendly platform. The enriched database Termi enables semantic enhancement and refinement of user's search query based on synonyms both in Serbian and German at a very high level. Serbian morphological e-dictionaries facilitate the morphological expansion of search queries in Serbian, thereby enabling the analysis of concepts and concept structures by identifying terms assigned to the concept, and by establishing relations between terms in Serbian and German which makes Bibliša a valuable Web tool that can support research and analysis of SrpNemKor.

Keywords Digital libraries, Aligned parallel corpora, Bilingual lexical resources, Lexical unit extraction, Bilingual search

Paper type Research paper

1. Introduction

Aligned multilingual corpora – bodies of text in parallel translation, also known as *bitexts* – have become an essential resource for work in multilingual natural language processing (NLP). These corpora represent the relationships between units in a source language and

The authors would like to thank Prof Dr Krstev Cvetana and Prof Dr Ranka Stanković for rereading carefully our paper several times and providing numerous suggestions for its improvement.

This research was supported by Serbian Ministry of Education and Science under the grants 178006 and TR 33003.

The Electronic Library
Vol. 37 No. 4, 2019
pp. 722-739
© Emerald Publishing Limited
0264-0473
DOI 10.1108/EL-03-2019-0056

their translation in a target language and, thus, they are an important digital resource for establishing equivalents between languages. More specifically, these corpora types allow researchers to determine the frequency of the occurrence of a particular word or a phrase defined in a search query in two or more languages, their grammatical forms and variants, as well as their semantic correlation with other words and phrases and their forms in two or more languages.

The Belgrade NLP group, researchers from the University of Belgrade of various backgrounds, has been developing language resources and tools for the processing of Serbian for decades, and some of which are the aligned multilingual corpora. Developed aligned texts are stored in the *Corpus of Contemporary Serbian* and in the aligned textual collections supported by the digital library Bibliša, both developed with the aim of enabling advanced search possibilities in the above-mentioned bilingual aligned textual collections. The *Corpus of Contemporary Serbian* (Utvić, 2013) contains several parallel corpora aligned with mostly literary texts but also texts from other domains, such as general news, scientific journals, Web journalism, health, law, education and movie subtitles[1].

The digital library Bibliša[2] contains collections of aligned texts from several scientific journals published bilingually in Serbia, texts produced within international projects, and some parallel corpora. This paper is focused on the Serbian-German literary corpus (SrpNemKor), which is stored in Bibliša. Furthermore, in the paper the possibilities of bilingual search of SrpNemKor in Bibliša are analysed based on existing bilingual Serbian-German lexical resources.

The next section presents a brief overview of previous work related to the bilingual terminology extraction based on parallel corpora. Section 3 provides the structure of the SrpNemKor and, subsequently, describes the pre-processing and alignment process of selected texts, as well as the language tools and resources used for these purposes. Section 4 introduces the web tool Bibliša, as well as the incorporation of SrpNemKor within this tool, while the search advantages of SrpNemKor within Bibliša followed by the evaluation of bilingual corpus search are described in Section 5. In Section 6, the achieved results and plans for further work are put forward.

2. Related work

2.1 Bilingual lexical extraction

Over the years, various researchers have used different techniques for multi-word unit (MWU) extraction and alignment to compile bilingual lexica. These approaches differ in terms of their methodology, used resources, languages involved and the purpose for which they have been built.

In several cases, the bilingual lists of MWUs were compiled to improve statistical machine translation of an existing machine translation system (Arcan *et al.*, 2017; Bouamor *et al.*, 2012; Irvine and Callison-Burch, 2016; Naguib, 2016; Oliver, 2017; Semmar, 2018; Tsvetkov and Wintner, 2010), for the development of an existing language resource in a target language on the basis of a corresponding resource in a source language – examples include the development of the Slovenian WordNet (Vintar and Fišer, 2008) based on the English WordNet and the development of the bilingual terminology based on the aligned corpora for the library and information science domain (Krstev *et al.*, 2018) – or for the presentation of bilingual correspondences between two languages – for example, correspondences between Slovak-Bulgarian parallel corpus (Garabík and Dimitrova, 2015).

Some of these approaches rely on the existence of a seed lexicon (Semmar, 2018; Tsvetkov and Wintner, 2010; Xu *et al.*, 2015) or existing translation memories and phrase tables (Oliver, 2017), while in some cases the existence of additional resources, in addition to the input corpus, is not required (Arcan *et al.*, 2017; Bouamor *et al.*, 2012; Garabík and Dimitrova, 2015; Naguib, 2016). Some approaches require parallel sentence-aligned data (Arcan *et al.*, 2017; Bouamor *et al.*, 2012; Garabík and Dimitrova, 2015; Semmar, 2018; Zhang and Wu, 2012), while others perform the extraction on comparable corpora (Hazem and Morin, 2016; Pinnis *et al.*, 2012; Xu *et al.*, 2015). The technique employed in Naguib (2016), used groups of aligned sentences (verses). In Irvine and Callison-Burch (2016), the authors performed two experiments, the first one relying on the existence of a bilingual dictionary with no parallel texts and the second one requiring only the existence of a small amount of parallel data. Bilingual lexica were compiled for different language pairs: English/French (Bouamor *et al.*, 2012; Hakami and Bollegala, 2017; Semmar, 2018), English/Spanish (Oliver, 2017), English/Arabic (Naguib, 2016), English/Italian and English/German (Arcan *et al.*, 2017), English/Slovene (Vintar and Fišer, 2008), English/Croatian, Latvian, Lithuanian and Romanian (Pinnis *et al.*, 2012), English/Chinese (Xu *et al.*, 2015; Zhang and Wu, 2012), English/Hebrew (Tsvetkov and Wintner, 2010), English/Italian (Arcan *et al.*, 2017), Slovak/Bulgarian (Garabík and Dimitrova, 2015), Serbian/English (Krstev *et al.*, 2018) and so on.

2.2 Bilingual corpus search

A digital library is a collection of information that is both digitized and organized (Lesk, 2005) and can be searched for different patterns (Bansode and Shinde, 2019; Bogaard *et al.*, 2019). Over the years, there has been a growing body of multilingual digital libraries (Diekema, 2012; Wu and Chen, 2019) that store content in more than one language. They provide various multilingual search query systems that offer users an efficient and useful way to exploit the collected parallel corpora retrieving parallel fragments that contain a word or a phrase from search query in any of the languages. Researchers (Gravano and Henzinger, 2014; Gutierrez-Vasques *et al.*, 2016; Volk *et al.*, 2014) argue that automatic word alignment and system and method using parallel corpora to translate terms from a search query from one to another language allows for major innovations in searching parallel corpora. This research relies on the existence of the parallel sentence-aligned Serbian-German literary corpus and the digital library Bibliša (Stanković *et al.*, 2017) that stores content in more than one language and provides multilingual keyword-based search invoking different translation lexical resources.

3. The Serbian-German aligned corpus

3.1 The content of the Serbian-German aligned corpus

The Serbian-German aligned literary corpus SrpNemKor is composed of contemporary novels in the Serbian and German languages. The first step in preparing this collection was to select text material sufficiently extensive and diverse, thereby enabling exploitation of developed language tools and technologies in the best possible way. Having investigated the availability of original texts and their translations, and having evaluated the quality of the available texts for processing, the researchers decided to include first only novels published in the second half of the twentieth century or the first half of the twenty-first century in this research phase. Also, the novels and authors awarded with some national literary award (such as the NIN Prize, Andrić Prize, etc., in Serbia) or an international literary award (such as the Nobel Prize) were selected. Finally, 14 novels were selected, these being seven novels written originally in Serbian with their German translations and seven novels written originally in German (four by Austrian authors and three by German authors) with their Serbian translations. According to this, SrpNemKor is divided in two sub-collections (Table I).

Table I. The list of selected novels for SrpNemKor

Authors	Texts in Serbian	Texts in German	English translation
Novels written originally in Serbian			
Albahari	Mamac	Mutterland	Bait
Arsenijević	U potpalublju: sapunska opera	Cloaca Maxima: eine Seifenoper	Cloaca Maxima : In the Hold : A Soap Opera
Kiš	Peščanik	Sanduhr	Hourglass
Olujić	Izlet u nebo	Ein Ausflug in den Himmel	An Excursion to the Sky
Tišma	Upotreba čoveka	Der Gebrauch des Menschen	The Use of Man
Valjarević	Komo	Como	Lake Como
Velikić	Ruski prozor	Das russische Fenster	*The Russian Window*
Novels written originally in German			
Bernhard	Moje nagrade	Meine Preise	My Prizes
Jelinek	Pijanistkinja	Die Klavierspielerin	Piano Teacher
Dor	Beč, juli 1999	Wien, Juli 1999	Vienna, July 1999
Grass	Hodom raka	Im Krebsgang	Crabwalk
De Bruyn	Buridanov magarac	Buridans Esel	Buridan's Ass
Ransmayr	Poslednji svet	Die letzte Welt	The Last World
Süskind	Parfem: Hronologija jednog zločina	Das Parfüm: die Geschichte eines Mörders	Perfume: The Story of Murderer

3.2 Text pre-processing and alignment

To create the corpus, all fourteen texts were pre-processed and aligned. Moreover, all texts were scanned, and afterwards OCR (optical character recognition) was applied. For the purpose of controlling the semi-automatic correction of texts in Serbian the electronic morphological dictionary of Serbian (Vitas and Krstev, 2006) was used, the only comprehensive electronic dictionary of Serbian that can be used in NLP, and the dedicated tool developed within Unitex[3] (Paumier, 2011). As a result of this process the existing e-dictionaries were enriched with almost 2,500 new entries. To control and correct texts in German Transkribus[4] was used, a platform for the automated recognition, transcription, and search of handwritten historical documents (Kahle *et al.*, 2017), as well as the language tool Hunspell[5], a spell checker and morphological analyser originally designed for the Hungarian language.

Before the alignment process, texts selected for SrpNemKor were transformed into XML documents and annotated to the sentence level. The annotation was done semi-automatically. However, sentence annotation was done automatically by means of Unitex and its finite-state transducers. Other texts' layouts (divisions, chapters, paragraphs) were annotated in two ways: with replacement method using regular expressions and manually. For the alignment process ACIDE (aligned corpora integrated development environment) was used, an integrated development environment for generating aligned parallel texts (Obradović *et al.*, 2008). More precisely, ACIDE enables automatic sentence alignment of texts with tool XAlign[6] and the alignment visualization and manual correction of alignment errors with the tool Concordancier[7], as well as automatic generation of documents in TMX format (translation memory eXchange)[8] (Savourel, 2004) requested for the corpus representation. During the alignment process of selected texts for SrpNemKor, 48,004 translation units (aligned sentences) were produced (Table II). Having completed the processing phase, the parallel corpus with fourteen novels containing more than 1.6 million words was prepared (Table III).

4. SrpNemKor in the digital library Bibliša

The parallel corpus SrpNemKor is stored as a resource of the web tool Bibliša. Bibliša has been developed within the Belgrade NLP group, aimed at enabling advanced search possibilities in multilingual digital libraries of e-journals, corpora and other text collections. It operates within a complex system comprised of several components: lexical resources (used to enhance and refine user queries), library content (documents in two languages aligned at the sentence level), Web services (used to access lexical resources) and a Web interface (used for user access to the library content) (Figure 1) (Stanković *et al.*, 2015).

All collections in Bibliša are cross-lingual searchable in a user-friendly environment in two ways: metadata search and full-text search by submitting keywords. Metadata search is monolingual and users need to select a language of search from a language list, then specify the scope of the search (all included text collections or only the specific ones) and then combine some of the metadata used for describing textual collections in Bibliša: words from a title, an author's name, a keyword, words from an abstract, and/or words from a document text. The obtained search results have the basic metadata description, such as document ID, author's name, title, links to all available metadata and a text in TMX with a limited number of translation units for unregistered users. For instance, for the query `Language: *SR* AND collection: *SrpNemKor* AND title: *prozor*", as a result a basic metadata description about the novel "Ruski prozor" (*The Russian Window*) in Serbian from the SrpNemKor collection was obtained as it contains "prozor" (window) in the title (Figure 2).

Table II.
The number of produced translation units in SrpNemKor by novels

Texts written originally in Serbian		Texts written originally in German	
Authors	No. of translation units	Authors	No. of translation units
Albahari	1,924	Bernhard	1,009
Arsenijević	1,174	de Bruyn	2,890
Kiš	4,112	Dor	1,249
Velikić	8,698	Gras	2,868
Valjarević	5,082	Jelinek	6,679
Tišma	3,937	Ransmayr	3,107
Olujić	1,361	Süskind	3,914
Total	26,288	Total	21,716
	48,004		

Table III.
The number of words in SrpNemKor by novels

Texts written originally in Serbian			Texts written originally in German		
Authors	Serbian	German		Serbian	German
Albahari	39,698	43,706	Bernhard	23,105	24,973
Arsenijević	24,329	25,194	de Bruyn	79,283	72,014
Kiš	64,571	68,131	Dor	24,300	23,136
Velikić	96,952	116,723	Gras	52,099	51,376
Valjarević	59,810	63,352	Jelinek	86,464	90,743
Tišma	91,217	97,116	Ransmayr	67,356	70,562
Olujić	29,117	28,210	Süskind	71,559	72,233
Total	405,694	442,432	Total	404,166	405,037
	848,126			809,203	
			1,657,329		

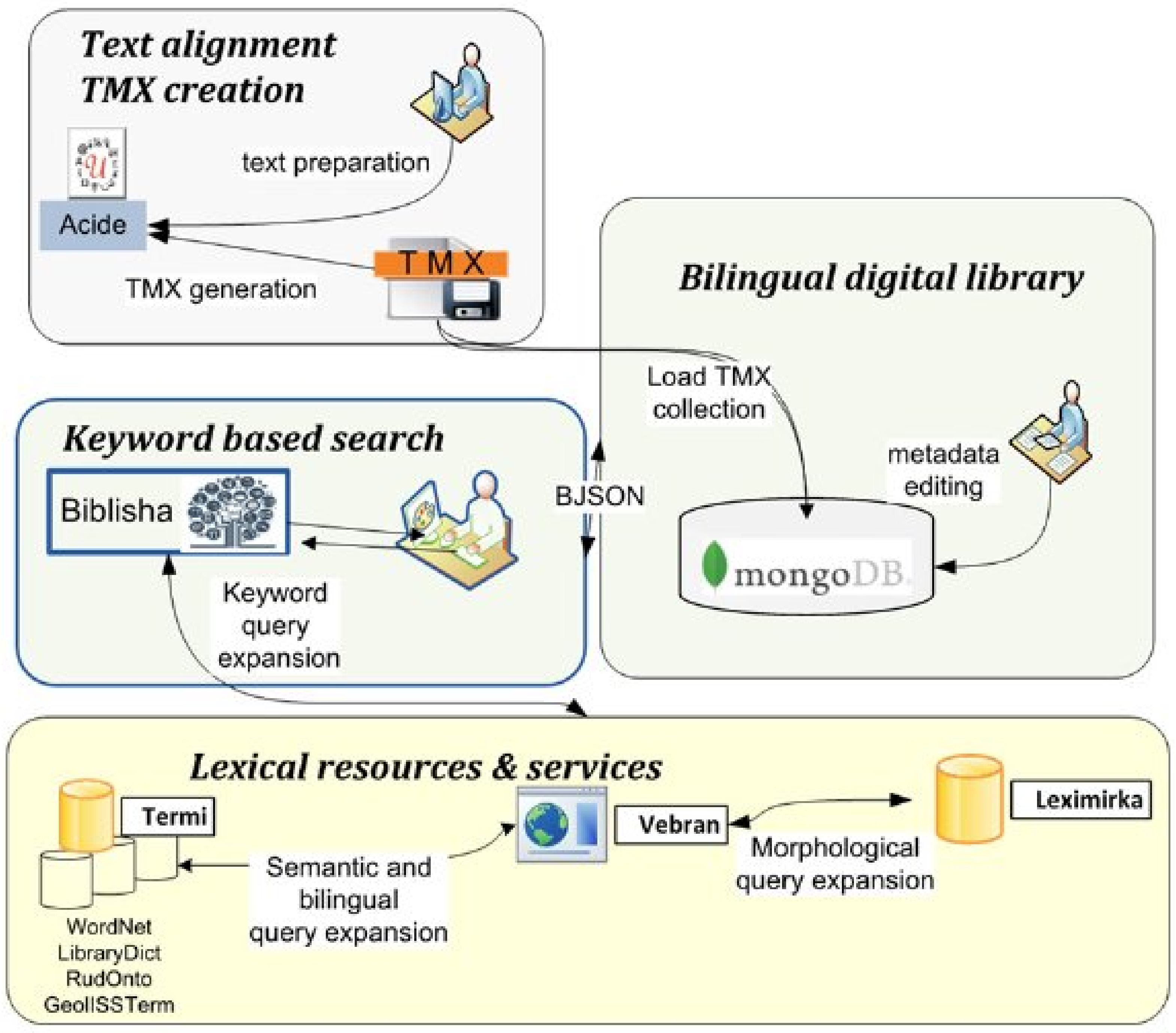


Figure 1.
Bibliša's system structure

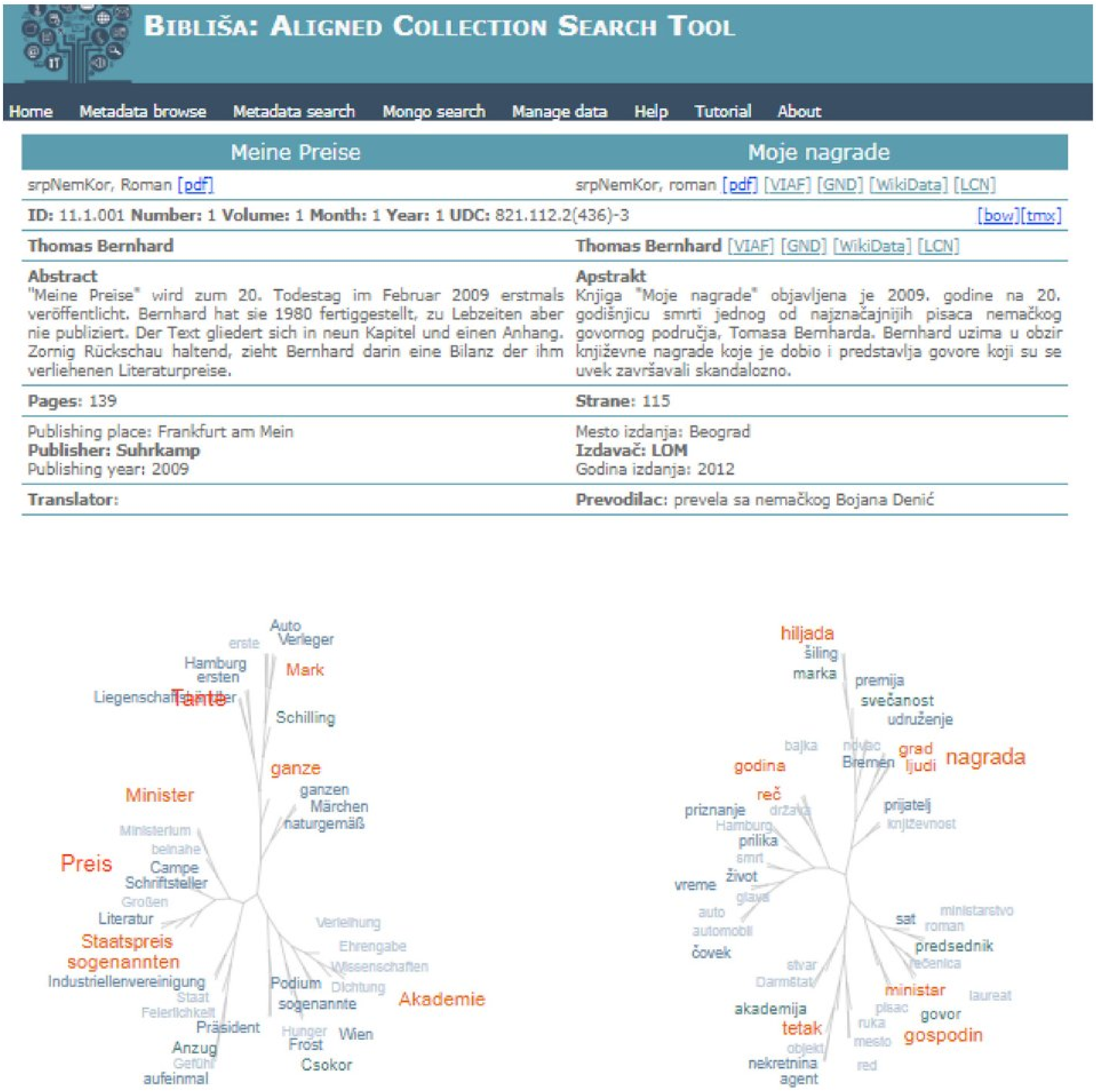
The main advantage of Bibliša is full text search and a possibility to enhance and refine users' bilingual queries both morphologically and semantically by using various lexical resources. The search results can be obtained as concordances of translation units that enable users to analyse all occurrences of a keyword from the query in the selected textual collection. It is explained in more detail in the next section.

5. Bilingual SrpNemKor search possibilities

5.1 Available lexical resources for bilingual search

In Bibliša's full-text search, a user has to specify first the scope of the search (all included text collections or only the specific one) and then formulate the initial query in the form of one or more simple or multi-word keywords which are then forwarded to the web service Vebran (Stanković *et al.*, 2011a). The system Vebran expands queries both semantically and morphologically (Stanković *et al.*, 2017) by invoking appropriate lexical resources: Serbian morphological electronic dictionaries, Serbian (Krstev *et al.*, 2004) and English WordNet (Fellbaum, 1998; Miller *et al.*, 1990), two domain terminological databases – the *Dictionary of Librarianship: Serbian-English and English-Serbian*[9] (Kovačević *et al.*, 2004) and the multilingual terminological database Termi[10] – and two ontologies, GeolISSTerm[11] (Stanković *et al.*, 2011b) and RudOnto[12] (Stanković *et al.*, 2014). All collections stored in Bibliša have been Serbian-English up to now. Consequently, all these resources enable expansion of search queries in Serbian and English. In searching SrpNemKor, in the first step, it was possible only to expand queries in Serbian by invoking Serbian morphological e-dictionaries and Serbian WordNet.

Serbian morphological electronic dictionaries are used for generating all inflected forms of query keywords in Serbian. These dictionaries were originally developed in the so-called DELA format with the direct influence and fruitful help of Maurice Gross and LADL (Laboratoire d'Automatique Documentaire et Linguistique). Today, they are incorporated into the Leximirka database (Stanković *et al.*, 2018) that also offers services for other NLP applications. The system of e-dictionaries containing general Serbian lexica (both for Cyrillic and Latin script) consists of a dictionary of simple word forms (a sequence of alphabetical characters), a dictionary of multi-word units (e.g. phrases and



BIBLIŠA: ALIGNED COLLECTION SEARCH TOOL

Meine Preise	Moje nagrade
srpNemKor, Roman [pdf]	srpNemKor, roman [pdf] [VIAF] [GND] [WikiData] [LCN]
ID: 11.1.001 **Number:** 1 **Volume:** 1 **Month:** 1 **Year:** 1 **UDC:** 821.112.2(436)-3	[bow][tmx]
Thomas Bernhard	**Thomas Bernhard** [VIAF] [GND] [WikiData] [LCN]
Abstract "Meine Preise" wird zum 20. Todestag im Februar 2009 erstmals veröffentlicht. Bernhard hat sie 1980 fertiggestellt, zu Lebzeiten aber nie publiziert. Der Text gliedert sich in neun Kapitel und einen Anhang. Zornig Rückschau haltend, zieht Bernhard darin eine Bilanz der ihm verliehenen Literaturpreise.	**Apstrakt** Knjiga "Moje nagrade" objavljena je 2009. godine na 20. godišnjicu smrti jednog od najznačajnijih pisaca nemačkog govornog područja, Tomasa Bernharda. Bernhard uzima u obzir književne nagrade koje je dobio i predstavlja govore koji su se uvek završavali skandalozno.
Pages: 139	**Strane:** 115
Publishing place: Frankfurt am Mein **Publisher: Suhrkamp** Publishing year: 2009	Mesto izdanja: Beograd **Izdavač: LOM** Godina izdanja: 2012
Translator:	**Prevodilac:** prevela sa nemačkog Bojana Denić

Figure 2.
Bibliša's metadata browser for the novel "*Meine Preise/Moje nagrade*" (*My Prizes*)

syntagma), and the set of finite-state transducers for the recognition of unknown words that are not recorded in the dictionary (Krstev, 2008; Krstev *et al.*, 2018; Vitas and Krstev, 2012b). Various special dictionaries, such as the dictionary of toponyms and other geopolitical names and the dictionary of Serbian personal names (Gucul-Milojević, 2010) as well as terminological dictionaries for various domains, have also been developed.

Serbian WordNet is a lexical semantic network based on Princeton WordNet's structure of concepts grouped into sets of cognitive synonyms (synsets). Its development was initiated in the scope of the BalkaNet WordNet project (Stamou *et al.*, 2002) and continues being developed ever since.

As it was necessary to enable expansion of queries in German, it was decided to update the lexical resource Termi. Termi is a multilingual database launched as a support for the development of terminological dictionaries in various domains (mathematics, computer science, mining, library science, computational linguistics, etc.). Until now, Termi has supported only the processing and representation of terms in Serbian and English. Based on the extraction of lexical units from the SrpNemKor aligned textual collection, Termi is enriched with new lexical units in Serbian and their German equivalents accompanied by their synonyms, thus enabling Bibliša to expand queries in German as well.

5.2 Enrichment of the database Termi

The list of lexical units added to Termi is extracted from the SrpNemKor aligned textual collection using the BilTE[13] system described in Krstev *et al.* (2018). It was originally designed for the generation of bilingual lexica of a domain multi-word terminology and it was applied to Serbian/English aligned domain corpora. This system was applied to the SrpNemKor corpus of literary novels, to compile a list of the most frequent translation equivalents occurring in the used corpus. In the following text, German is considered to be a source language, and Serbian is a target language. BilTE operates on the assumption that if a list of terms, in this case lexical units, exists for a source language, as well as an aligned corpus for a source and a target language, and an extractor of MWUs for a target language, then it is possible to extract the list of corresponding term translations for a target language. This assumption was confirmed for the Serbian/English language pair and a domain corpus. For the present project, the following resources were used:

- Aligned texts: The obtained Serbian-German literary textual collection containing 48,004 aligned sentences;
- A list of lexical units for a source language: this list of lexical units has been prepared by means of two sources. The first one is Open-DE-WordNet (Open DE WordNet Initiative, 2019), open German WordNet, with approximately 120,000 units, while the second one is a list of the most frequent German words available in Wiktionary: Frequency lists[14]. The list approximately includes 10,000 units. To achieve better results, the authors in Krstev *et al.* (2018) performed a word-by-word lemmatization on a source list of lexical units. The German model available in spaCy[15] was used that performs lemmatization using an underlying lookup table of 355,354 single-word units. After lemmatization and duplicates elimination, the final list contains 27,638 distinct German lexical units; and
- An extractor of multi-word lexical units for a target language: LeXimir, a system for multi-word lexical units' extraction for Serbian, based on e-dictionaries and local grammars described in Stanković *et al.* (2016) is used. Since this system extracts multi-word lexical units, a list of single-word lexical units from the merged Serbian texts from SrpNemKor using Unitex was separately prepared. The obtained list contained 94,802 single-word and 48,159 multi-word lexical units. Afterwards, lemmatization was performed using e-dictionaries for Serbian (see Section 5.1). After lemmatization, duplicates were eliminated, yielding the total of 77,297 distinct Serbian lexical units.

The BilTE system receives as an input the following: aligned texts, list of lemmatized lexical units for a source language, as well as the list of lemmatized lexical units for a target language. Subsequently, it applies text processing (cleaning, tokenization, truecasing) of the aligned texts, after which word-alignment, phrase extraction and phrase scoring are performed, and lexicalized reordering tables are created using GIZA++ (Koehn *et al.*, 2003; Och and Ney, 2003). These so-called "phrase tables" contain aligned chunks in German and Serbian that represent potential translation pairs. The first step afterwards is cleaning; that is, removal of various character entities or chunks that consist only of digits and/or punctuation. In addition to bilingual chunks of text, the phrase table includes, among others, a probability of the source chunk being translated as a target chunk and vice versa. If any of these two values is less than 0.85, the chunk is discarded by the system. The third filtering step in this phase is a removal of chunks that definitely do not belong to the list of lexical units for the source language. For this purpose, the list of lexical units was represented as a union of distinct tokens (i.e. bag-of-words (BoW) representation was obtained). The system then checks whether a source chunk contains at least one token present in the BoW. If there is no intersection with the BoW representation of the lexical units list, the chunk is also discarded.

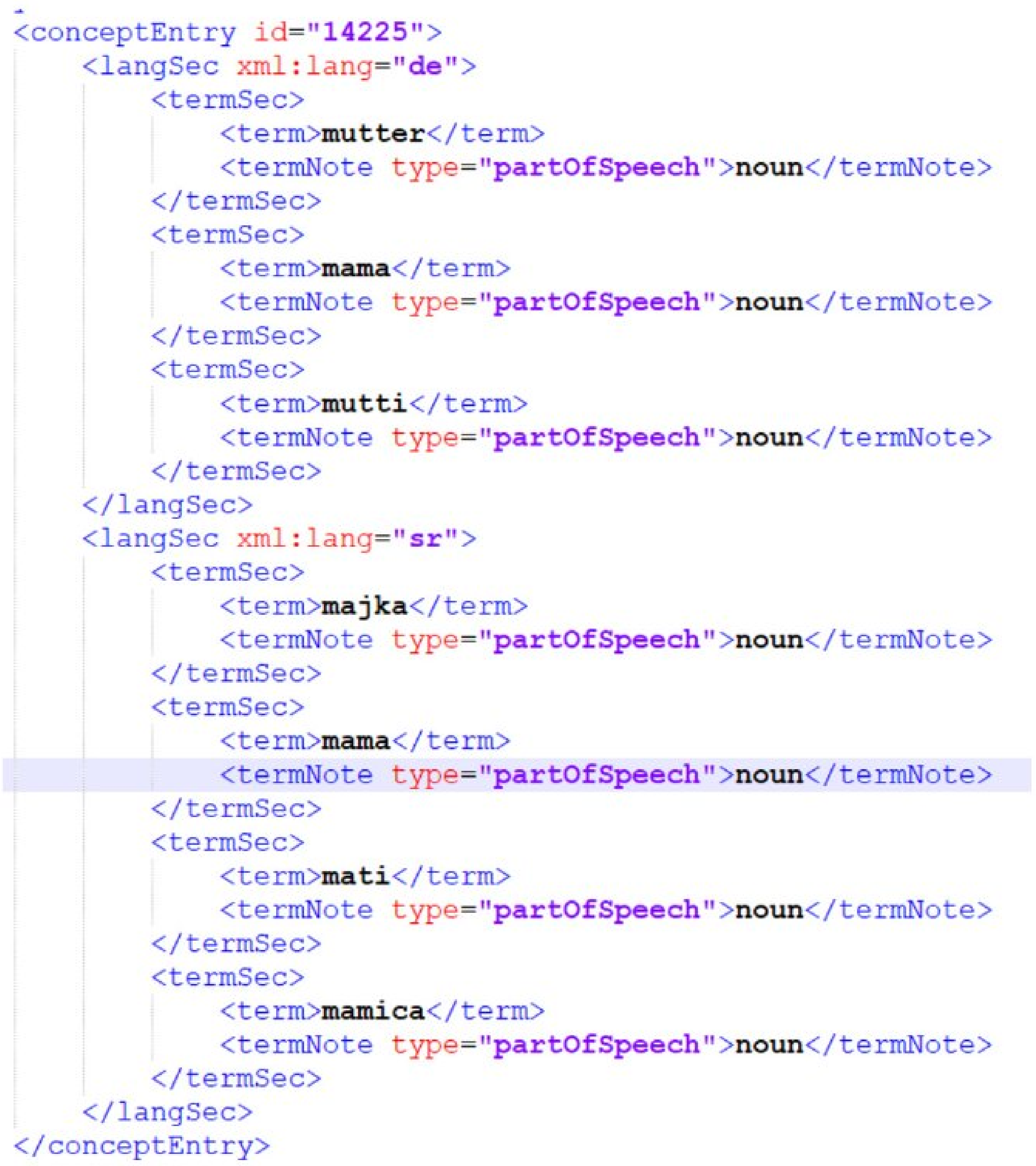
In the first phase, only pairs from the phrase table having a German chunk that matched some lexical units from the previously prepared list of German lexical units were kept. This "match" relation is defined as follows: if a chunk is represented by an unordered set of distinct words obtained from it after removal of stop-words and lemmatization, the two strings match if they are represented by the same set. For example, if a lexical unit is "natural language processing" and a chunk is "the processing of natural languages", their corresponding set representations (after lemmatization and stop-words removal) are {natural, language, processing} and {processing, natural, language}, respectively. Since these two sets are equal, it is considered that they match. This is used for both sides of the aligned corpus. In the second phase, pairs with a Serbian chunk that were not matched by any lexical unit from the extracted list were eliminated from the final list.

After applying the BilTE system and tool GIZA++ to the resources and additional filtering, a list of 14,142 candidate translation pairs was obtained. One of the authors manually evaluated the obtained list of translated pairs. Candidates were removed from the list for several reasons: (a) a Serbian and a German part in a pair did not match, for example "kap po kap" (*drop by drop*)/"heraus" (*outside*) or "Karlovom primeru" (*Karl's example*)/"sehen" (*to see*); (b) a Serbian part or a German part in a pair did not represent a reasonable phrase, for example "deo mraz *s* naslov" (*no sense*)/Frost (*frost*); (c) a Serbian part is longer than expected from a German counterpart, for example "komad papira" (*a piece of paper*)/"papier" (*paper*) or vice versa, a German part is longer, for example "kafa" (*coffee*)/"kaffe trinken" (*to drink coffee*). After evaluation, a final list of 3,984 correct translation pairs was obtained. Some examples of pairs from the list are given in Table IV, while the list of all pairs manually evaluated as correct is available online[16].

The evaluated translation pairs were further analysed, synonym candidates identified and imported into the Termi database[17]. For example, after further analyses it was identified that the synonym "Mutti" for "Mutter" was not in the list of German lexical units and it is added in Termi. The authorized user of the Termi application can export selected datasets in various formats: Excel, CSV, TBX[18] (Figure 3).

Table IV.
Some aligned German-Serbian lexical units from the SrpNemKor

Target term	Source term	English
autobuskoj stanici	Haltestelle	*Bus station*
baterijska lampa	Taschenlampe	*Flashlight*
beznadežan slučaj	Hoffnungslos Fall	*Hopeless case*
aleja	allee	*Alley*
starica	alt Frau	*Old woman*
omladinski dom	Jugendherberge	*Youth hostels*
pokretom ruke	Handbewegung	*Hand movement*
boravišna dozvola	Aufenthaltsgenehmigung	*Residence permit*



```
<conceptEntry id="14225">
    <langSec xml:lang="de">
        <termSec>
            <term>mutter</term>
            <termNote type="partOfSpeech">noun</termNote>
        </termSec>
        <termSec>
            <term>mama</term>
            <termNote type="partOfSpeech">noun</termNote>
        </termSec>
        <termSec>
            <term>mutti</term>
            <termNote type="partOfSpeech">noun</termNote>
        </termSec>
    </langSec>
    <langSec xml:lang="sr">
        <termSec>
            <term>majka</term>
            <termNote type="partOfSpeech">noun</termNote>
        </termSec>
        <termSec>
            <term>mama</term>
            <termNote type="partOfSpeech">noun</termNote>
        </termSec>
        <termSec>
            <term>mati</term>
            <termNote type="partOfSpeech">noun</termNote>
        </termSec>
        <termSec>
            <term>mamica</term>
            <termNote type="partOfSpeech">noun</termNote>
        </termSec>
    </langSec>
</conceptEntry>
```

Figure 3.
An example TBX format for term "Mutter" (mother) exported from Termi

5.3 An example of SrpNemKor full-text search

As stated in Section 5.1, SrpNemKor searching of Bibliša enables semantic and morphological expansion of search queries by invoking Serbian morphological e-dictionaries, Serbian WordNet and Termi. By checking the option "Morphological query expansion" in the search browser, Bibliša enables morphological expansion of queries in Serbian and users retrieve concordances containing all inflected forms of each query keyword found in the SrpNemKor collection. For example, without morphological expansion only text segments containing the basic form "mama" (in Serbian) would be retrieved, while with this option all text segments containing some singular form "mame", "mami", "mamu", "mamom" or plural form "mame", "mamama" would be retrieved as well. Another example of a keyword search query with a multi-word unit is presented below. For example, without morphological expansion only text segments containing the basic form "muzički instrument" (in Serbian) would be retrieved, while with this option included, all text segments containing some singular form "muzičkom instrumentu", "muzičkog instrumenta", or plural form "muzički instrumenti", "muzičkih instrumenata" would be retrieved as well. At this moment, the morphological expansion of queries in German is not provided because an appropriate open resource that could be used for this purpose was not found.

For semantic expansion of queries in Serbian Bibliša invokes the Serbian WordNet. For example, for the query "Language: *SR* AND collection: *SrpNemKor* AND keyword: *nebo*", the Serbian WordNet expands the query with the lexical units: "Eldorado, nebesa, nebeska sfera, nebeski svod, nebo, nirvana, obećana zemlja, raj". During this research, some possible open source German WordNet for semantic expansion of queries in German were analysed, but they are not yet sufficiently developed so neither of them were included.

For additional expansion of queries in searching SrpNemKor, Bibliša invokes the updated lexical resource Termi. For instance, as a result of the query:

Q1. "Language: *SR* and collection: *SrpNemKor* and keyword: *majka*"

several terms in both Serbian and German are obtained as synonyms of the keyword "majka" (mother). In this particular example, German terms expanding the query are: "Mama", "Mutter" and "Mutti" while the corresponding Serbian terms are: "majka", "mama", "mati" and "mamica". The obtained additional keywords can be used without modification or a user can change them by removing some keywords or by adding new ones.

A user can further specify search by choosing one of the following options: "DE&SR", "DE" and "SR". When the first one "DE&SR" is checked, keywords retrieved in both languages, in the present case in German and Serbian, are highlighted in the generated aligned concordances. For the given query (Q1), 1,413 aligned concordances were retrieved. The full list can be seen only by registered users; unregistered users can obtain up to ten randomly chosen lines. Four of these lines are shown in Table V.

When options "DE" or "SR" are checked, the generated aligned concordances have the keywords matched only in the German (140 concordances) or only in the Serbian part (67 concordances) of the aligned corpus, respectively. The resulting concordances in which only German keywords are retrieved and highlighted are shown in Table VI. Serbian terms were

Table V.
Produced concordances for the query "majka" (mother) using the search option "DE&SR"

Ruski prozor = Das russische Fenster/Dragan Velikić, ID: 11.2.007metadata	n1023 Es kam vor, dass er ins Stadtzentrum ging, nur um zu überprüfen, ob noch immer der Film gezeigt wurde, den *Mama* und er so gerne sehen wollten	n1023 Dešavalo se da ode u centar grada samo da bi proverio da li je i dalje na repertoar u film koji su *majka* i on želeli da vide
Ruski prozor = Das russische Fenster/Dragan Velikić, ID: 11.2.007metadata	n2904 Darüber dachte er nach, wenn er sich nach den Telefonaten mit seiner *Mutter* in sein Lieblingscafe in Budim begab	n2904 O tome razmišlja kada posle subotnjih razgovora sa *majkom* krene u omiljeni kafe na Budimu
Upotreba čoveka = Der Brauch des Menschen/ Aleksandar Tišma, ID: 11.2.005metadata	n1686 Die *Mutter* hat Vera schon am Tag nach ihrer Ankunft ermahnt: »Kein Wort über Vater und Gerd, das würde keiner verstehen	n1686 *Mati* je Veru već sutradan po dolasku upozorila: „Nemoj im ništagovoriti o ocu i Gerdu, oni to ne bi razumeli
Die Klavierspielerin = Pijanistkinja/Elfride Jelinek, ID: 11.1.002metadata	n210 Die *Mutter* soll streng ihr Gewissen erforschen, ob sie ein ähnlich geschnittenes Kleid nicht in ihrer Jugend selbst getragen habe, *Mutti*?	n210 Neka *majka* strogo ispita svoju savest, nije li ona sama u svojoj mladosti nosila neku slično krojenu haljinu, *mamice*?

Table VI.
Produced concordances for the query "majka" (mother) using the search option "DE"

Mamac = Mutterland/ David Albahari, ID: 11.2.001metadata	n196 Vielleicht ist meine Äußerung über die politische Einstellung des ersten Mannes meiner *Mutter* ungerecht	n196 Možda grešim kada govorim o političkim opredeljenjima majčinog prvog muža
Mamac = Mutterland/ David Albahari, ID: 11.2.001metadata	n414 *Mutter* war ein Traum gewesen, der in einem fremden Traum gelebt hatte	n414 bila je san koji je živeo u tuđem snu

not retrieved due to different reasons. In the first line, the Serbian translation equivalent "majčinog" was not recognized as it is an adjective derived from "majka", however, it is not in the synonymy relation with the keyword (noun) "majka". In the second line, the author did not use "majka" while the translator used "Mutter". Table VII illustrates three aligned concordance lines with only Serbian terms retrieved and highlighted. In the first two lines, translators did not use a German equivalent for the keyword "majka", or more precisely for forms "majci" (the dative case singular) and "majke" (the genitive case singular), because translators used "Eltern" (parents) and "splitternackt" (stark naked) for Serbian phrases "ocu i majci" (father and mother) and "gola kao od majke rođena" (naked as her mother bore her). In the last concordance line, the German term "Mütter" is not highlighted as it is not available in Termi in this particular form. The term "Mütter" is a plural from of "Mutter" and it would have been recognized if the morphological expansion of German keywords could have been provided. Presently, it is added as a new entry for Termi.

For each concordance line, there is an identification of the document it originates from containing a link to the full metadata in both languages of a retrieved document (Tables V to VII).

5.4 The evaluation of bilingual search for SrpNemKor

To evaluate the bilingual search of SrpNemKor in the Bibliša digital library, the following ten examples of lexical units in Serbian were selected (their intended meanings are indicated in English): mladić (*young man*), zvuk (*sound*), pisac (*writer*), stanica (*station*), osmeh (*smile*), čas (*moment*), potvrda (*certificate*), jak (*strong*), glavni grad (*capital city*) and kupiti (*to buy*). All lexical units were selected from the table of translated pairs analysed in Section 5.2. To evaluate the system behaviour in the best possible way, selected for queries were lexical units that have synonyms in at least one of the languages. For example, in Serbian mladić/Junge has a synonym momak, while the lexical unit stanica/Haltestelle has synonyms both in Serbian stajalište and German Station. The evaluation comprised of the analysis of the obtained search results for five query types for all selected lexical units:

(1) a basic query without morphological, semantic (synonyms) and language expansions (o);

(2) a query with morphological expansions for Serbian (m);

(3) a query with semantic expansions for Serbian (s);

(4) a query with morphological and semantic expansions for Serbian (ms); and

(5) a query that includes equivalents in German with all expansions in both languages (msde).

Cloaca Maxima = Cloaca Maximar/Vladimir Arsenijević, ID: 11.2.002metadata	n661 Für meine Eltern – wie für viele Städter – war das Landleben ein Quell exotischer Freuden	n661 Kao mnogim ljudima koji su život proveli u gradu, mom ocu i *majci* selo predstavlja izvor egzotičnih uživanja
Peščanik = Sanduhr/Danilo Kiš, ID: 11.2.003metadata	n3860 Eva, splitternackt, hat mit der Rechten den untersten Ast ergriffen, während sie zwischen den Fingern der Linken den Apfel hält, den sie Adam anbietet	n3860 Eva, gola kao od *majke*, uhvatila je desnom rukom najnižu granu, a u levoj ruci, između stisnutih prstiju drži jabuku, pružajući je Adamu
Burdans Esel = Buridanov magarac/Ginter de Brojn, ID: 11.1.005metadata	n635 Selbst feinsinnigste Mädchen lernen als Hausfrauen und Mütter rechnen	n635 Čak i najprefinjenije devojke nauče da računaju kad postanu domaćice i *majke*

Table VII. Produced concordances for the query "majka" (mother) using the search option "SR"

Three persons manually evaluated the obtained search results both in the Serbian and the German part of the corpus, as well as in the Serbian-German translated pairs. For each type of the search query and each lexical unit precision (P) and recall (Q) were calculated: the precision represents the ratio of relevant retrieved instances and all retrieved instances for each query; the recall represents the ratio of relevant retrieved instances and the total amount of relevant instances retrieved by the widest query (presuming that this set is the closest to the set of all relevant instances in the corpus; that is why the Qmsde is 1.00 in Figure 5). The F1 measure is the harmonic mean of the precision and the recall.

The analysis of the obtained results in Figure 5 shows that in most cases, morphological expansion contributes more to the increase of the recall than semantic expansion (see queries for mladić and osmeh, the only exceptions are queries for potvrda and zvuk). However, the query glavni grad shows significant increase in the recall with morphological and semantic expansions because its synonym in Serbian prestonica "appeared to be more frequent in SrpNemKor.

The results presented in Figure 4 confirm the expected decrease of the precision with the enhancement of the queries. The morphological expansion can introduce homographic forms: for instance, one of the inflected forms of the query term kupiti is *kupe* (the third person singular of the present tense) that coincides with the noun *kupe* (compartment) in the nominative singular and the noun *kupa* (cone) in the genitive singular. The semantic expansion can also lead to the decrease of the precision because of the polysemy of the query lexical unit. For instance, the query čas, in the present case meaning a moment, yielded results with different meanings of the same word: an hour and a school class.

However, one can observe in Figure 6 the search improvement for all query expansions compared to the basic query (Fo). The query that includes all expansions performs best (Fmsde) except in one case: the German equivalent of mladić is the noun Junge that is homographous with some forms of the frequent adjective *junge* that decreased the precision.

6. Conclusion

In this paper, the researchers analysed one parallel literary corpus with the aligned Serbian-German texts and the ways in which it can be searchable in one complex digital library environment as Bibliša. Also, the importance of the bilingual lexical resources in searching this kind of aligned textual collection has been presented as well as the manner in which lexical resources, such as database Termi, can be enriched with translated pairs of lexical

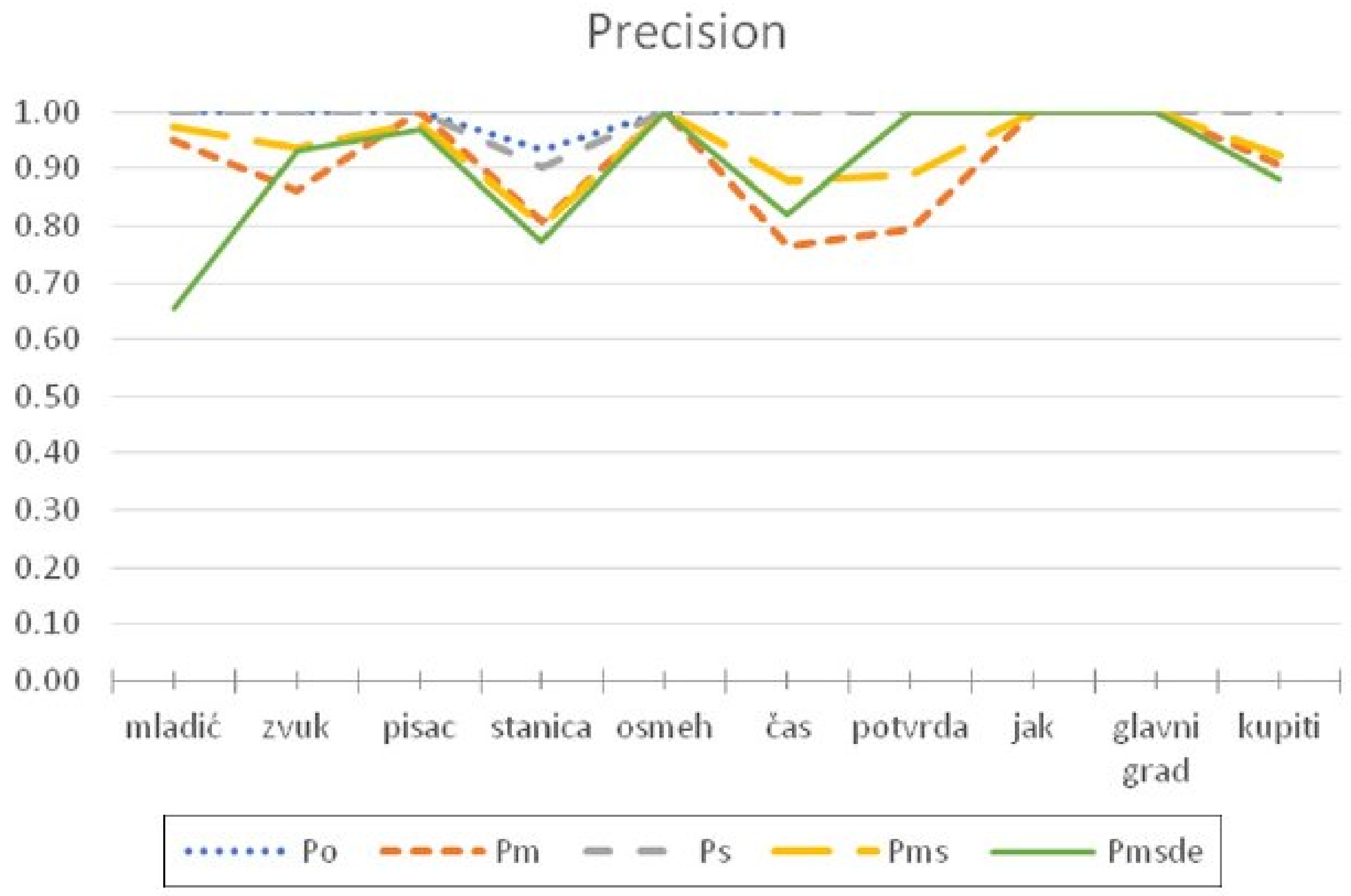


Figure 4.
The precision for search queries in SrpNemKor

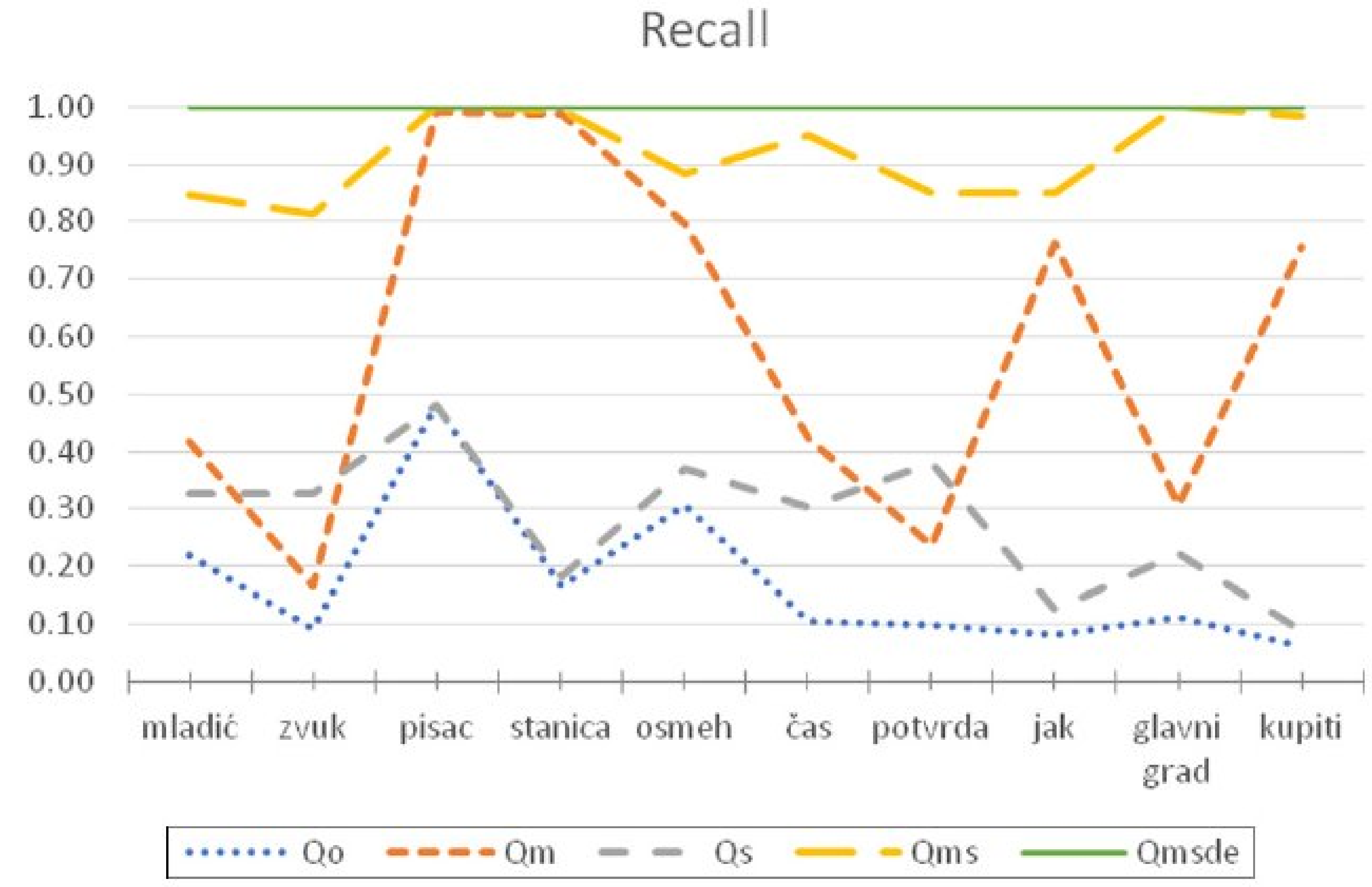


Figure 5. The recall for search queries in SrpNemKor

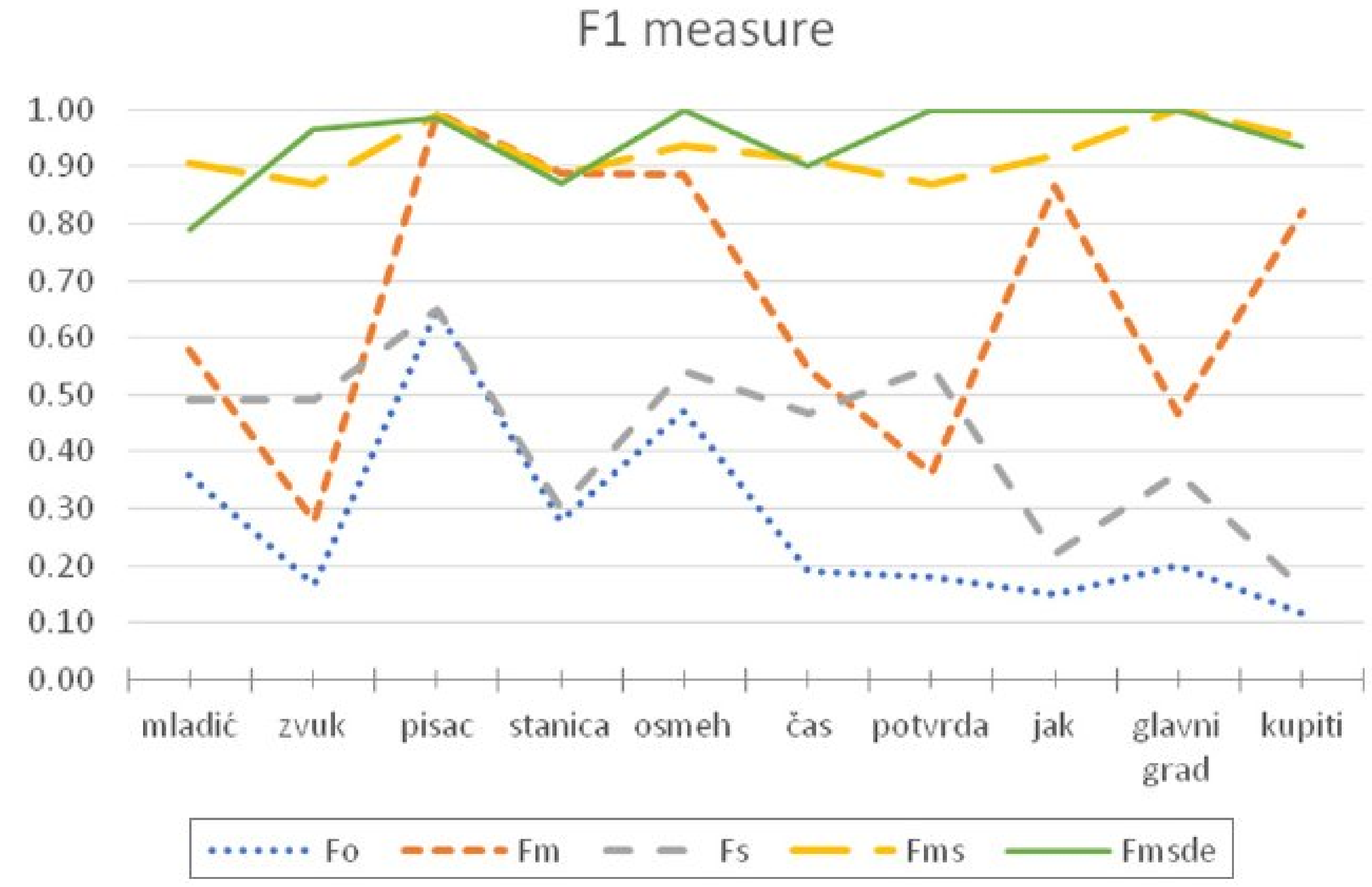


Figure 6. F1 measure for search queries in SrpNemKor

units extracted from a parallel literary corpus aimed at enabling the enhancement of search queries.

In the future, there is a plan to enhance and refine the SrpNemKor search capabilities. One of the goals would be to add more German lexical resources to Bibliša, primarily a morphological generator and a WordNet to improve morphological and semantic expansion of search queries in German. A further goal is to annotate the named entities in SrpNemKor and to produce the mapping of the annotated named entities between languages that would improve their recognition and enrich annotations on both the German and Serbian sides.

Notes

1. Parallel corpora in the *Corpus of Contemporary Serbian* can be searched with a personal user account. All of them are available at: http://poincare.matf.bg.ac.rs/~cvetana/LT-pregled-en.html with official data about their size.
2. Bibliša is available at: http://jerteh.rs/biblisha/

3. Unitex, the open source multilingual corpus processing suite, is available at: https://unitexgramlab.org
4. Transkribus is available at: https://transkribus.eu/Transkribus/
5. Hunspell is available at: http://hunspell.github.io
6. Xalign is available at: https://github.com/qfish/XAlign
7. Concordancier is available at: www.univ-montp3.fr/sl/rachel/E42SLL1/concordancier.htm
8. TMX is an open XML-based standard intended for easier exchange of translation memories of data that is aligned parallel texts, between tools and translation vendors. TMX is available at: http://xml.coverpages.org/tmxSpec971212.html
9. The *Dictionary of Library and Information Science* is available at: http://rbi.nb.rs/en/dict.html
10. Termi is available at: http://termi.rgf.bg.ac.rs
11. GeolISSTerm (Geologic Information System of Serbia) is available at: http://geoliss.mre.gov.rs/recnik/
12. RudOnto is available at: http://rudonto.rgf.bg.ac.rs
13. Bilingual Terminology Extraction tool is available at: http://bilte.jerteh.rs/
14. Wiktionary: Frequency lists for the German language are available at: https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#German
15. spaCyLemmatizer is available at: https://spacy.io/api/lemmatizer
16. The obtained list of translation pairs is available at: http://portal.jerteh.rs/tematres
17. A list of translated pairs of lexical units extracted from SrpNemKors is available at: http://termi.rgf.bg.ac.rs/Dodaj/Dodaj/111929
18. Term Base eXchange is an international standard defined by ISO (30042), initially by Localization Industry Standards Association (LISA). It defines an XML format for interchange (exchange) of terminological data between different terminological databases. It is available at: www.iso.org/standard/62510.html

References

Arcan, M., Turchi, M., Tonelli, S. and Buitelaar, P. (2017), "Leveraging bilingual terminology to improve machine translation in a computer aided translation environment", *Natural Language Engineering*, Vol. 23 No. 5, pp. 763-788.

Bansode, N.N. and Shinde, M.G. (2019), "Impact of new technologies in the digital libraries", *Journal of Advancements in Library Sciences*, Vol. 6 No. 1, pp. 279-283.

Bogaard, T., Hollink, L., Wielemaker, J., van Ossenbruggen, J. and Hardman, L. (2019), "Metadata categorization for identifying search patterns in a digital library", *Journal of Documentation*, Vol. 75 No. 2, pp. 270-286.

Bouamor, D., Semmar, N. and Zweigenbaum, P. (2012), "Identifying bilingual multi-word expressions for statistical machine translation", in Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. and Piperidis, S. (Eds), *Proceedings of LREC'12 Conference*, pp. 674-679.

Diekema, A.R. (2012), "Multilinguality in the digital library: a review", *The Electronic Library*, Vol. 30 No. 2, pp. 165-181.

Fellbaum, C. (Ed.) (1998), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge.

Garabík, R. and Dimitrova, L. (2015), "Extraction and presentation of bilingual correspondences from Slovak-Bulgarian parallel corpus: cognitive studies", *Études Cognitives*, Vol. 15, pp. 327-334.

Gravano, L. and Henzinger, M.H. (2014), "Systems and methods for using anchor text as parallel corpora for cross-language information retrieval", US Patent 8,631,010.

Gucul-Milojević, S. (2010), "Personal names in information extraction", *Infotheca*, Vol. 11 No. 1, pp. 53a-63a.

Gutierrez-Vasques, X., Sierra, G. and Pompa, I.H. (2016), "Axolotl: a web accessible parallel corpus for Spanish-Nahuatl", *Proceedings of LREC'16 Conference, European Language Resources Association*, pp. 4210-4214.

Hakami, H. and Bollegala, D. (2017), "A classification approach for detecting crosslingual biomedical term translations", *Natural Language Engineering*, Vol. 23 No. 1, pp. 31-51.

Hazem, A. and Morin, E. (2016), "Efficient data selection for bilingual terminology extraction from comparable corpora", *Proceedings of COLING'16*, pp. 3401-3411.

Irvine, A. and Callison-Burch, C. (2016), "End-to-end statistical machine translation with zero or small parallel texts", *Natural Language Engineering*, Vol. 22 No. 4, pp. 517-548.

Kahle, P., Colutto, S., Hackl, G. and Mühlberger, G. (2017), "Transkribus: a service platform for transcription, recognition and retrieval of historical documents", *14th IAPR International Conference on Document Analysis and Recognition (ICDAR'17)*, Vol. 4, *IEEE*, pp. 19-24.

Koehn, P., Och, F.J. and Marcu, D. (2003), "Statistical phrase-based translation", *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Vol. 1, *Association for Computational Linguistics*, pp. 48-54.

Kovačević, L.J., Injac, V. and Begenišić, D. (2004), *Bibliotekarski Terminološki Rečnik: Englesko-Srpski, Srpsko-Engleski (Library Terminological Dictionary: English-Serbian, Serbian-English)*, Narodna biblioteka Srbije, Beograd.

Krstev, C. (2008), *Processing of Serbian – Automata, Texts and Electronic Dictionaries, Faculty of Philology*, University of Belgrade, Belgrade.

Krstev, C., Pavlović-Lažetić, G., Vitas, D. and Obradović, I. (2004), "Using textual and lexical resources in developing Serbian WordNet", *Romanian Journal of Information Science and Technology*, Vol. 7 Nos 1/2, pp. 147-161.

Krstev, C., Šandrih, B., Stnaković, R. and Mladenović, M. (2018), "Using English baits to catch Serbian multi-word terminology", in Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. and Tokunaga, T. (Eds), *Proceedings of LREC'18 Conference*, pp. 2487-2494.

Lesk, M. (2005), *Understanding Digital Libraries*, Elsevier, San Francisco, CA.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J. (1990), "Introduction to WordNet: an on-line lexical database", *International Journal of Lexicography*, Vol. 3 No. 4, pp. 235-244.

Naguib, S.Y.M. (2016), "Bilingual lexicon extraction from Arabic-English parallel corpora with a view to machine translation", *Arab World English Journal*, Vol. 7 No. 5, pp. 317-336.

Obradović, I., Stanković, R. and Utvić, M. (2008), "An integrated environment for development of parallel corpora (in Serbian)", *Die Unterschiede Zwischen Dem Bosnischen/Bosniakischen, Kroatischen Und Serbischen*, LitVerlag, Münster, pp. 563-578.

Och, F.J. and Ney, H. (2003), "A systematic comparison of various statistical alignment models", *Computational Linguistics*, Vol. 29 No. 1, pp. 19-51.

Oliver, A. (2017), "A system for terminology extraction and translation equivalent detection in real time: efficient use of statistical machine translation phrase tables", *Machine Translation*, Vol. 31 No. 3, pp. 147-161.

Open DE WordNet Initiative (2019), available at: https://ikum.mediencampus.h-da.de/projekt/open-de-wordnet-initiative/ (accessed 28 February 2019).

Paumier, S. (2011), "Unitex 3.0 user manual", available at: www.cis.uni-muenchen.de/people/lg3/ManuelUnitex.pdf (accessed 15 December 2018).

Pinnis, M., Ljubešic, N., Stefanescu, D., Skadina, I., Tadić, M. and Gornostay, T. (2012), "Term extraction, tagging, and mapping tools for under-resourced languages", *Proceedings of TKE'12*, pp. 20-21.

Savourel, Y. (2004), "TMX 1.4b specification", available at: www.gala-global.org/tmx-14b (accessed 25 December 2018).

Semmar, N. (2018), "A hybrid approach for automatic extraction of bilingual multiword expressions from parallel corpora", in Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. and Tokunaga, T. (Eds), *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*, *European Language Resources Association*, *Paris*.

Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D. and Grigoriadou, M. (2002), "Balkanet: a multilingual semantic network for the Balkan languages", *Proceedings of the International WordNet Conference*, *Mysore*, pp. 21-25.

Stanković, R., Obradović, I. and Utvić, M. (2014), "Developing termbases for expert terminology under the TBX standard", in Pavlović-Lažetić, G., Krstev, C., Obradović, I. and Vitas, D. (Eds), *Natural Language Processing for Serbian: Resources and Applications*, University of Belgrade, Faculty of Mathematics, Belgrade, pp. 12-26.

Stanković, R., Krstev, C., Lazić, B. and Vorkapić, D. (2015), "A bilingual digital library for academic and entrepreneurial knowledge management", in Spender, J.C., Schiuma, G. and Albino, V. (Eds), *Proceeding of IFKAD'15: Culture, Innovation and Entrepreneurship: Connecting the Knowledge Dots*, pp. 1764-1777.

Stanković, R., Krstev, C., Lazić, B. and Škorić, M. (2018), "Electronic dictionaries – from file system to lemon based lexical database", in McCrae, J.P., Chiarcos, C., Declerck, T., Gracia, J. and Klimek, B. (Eds), *Proceedings of LREC'18 – W23 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science (LDL'18)*, *European Language Resources Association*, *Paris*, pp. 48-56.

Stanković, R., Krstev, C., Obradović, I., Lazić, B. and Trtovac, A. (2016), "Rule-based automatic multi-word term extraction and lemmatization", in Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S. (Eds), *Proceedings of LREC'16*, *European Language Resources Association*, *Paris*, pp. 507-514.

Stanković, R., Obradović, I., Krstev, C. and Vitas, D. (2011a), "Production of morphological dictionaries of multi-word units using a multipurpose tool", in Jassem, K., Fuglewicz, P.W., Piasecki, M. and Przepiórkowski, A. (Eds), *Proceedings of the Computational Linguistics-Applications Conference, 17-19 October*, *Polish Information Processing Society*, *Poland*, pp. 77-84.

Stanković, R., Krstev, C., Vitas, D., Vulović, N. and Kitanović, O. (2017), "Keyword-based search on bilingual digital libraries", in Calì, A., Gorgan, D. and Ugarte, M. (Eds), *Semantic Keyword-Based Search on Structured Data Sources – Second COST Action IC1302 International KEYSTONE Conference, IKC'16*, *Springer*, pp. 112-123.

Stanković, R., Trivić, B., Kitanović, O., Blagojević, B. and Nikolić, V. (2011b), "The development of the GeolISSTerm terminological dictionary", *INFOtheca*, Vol. 12 No. 1, pp. 49a-63a.

Tsvetkov, Y. and Wintner, S. (2010), "Extraction of multi-word expressions from small parallel corpora", *Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING'10*, Association for Computational Linguistics, *Stroudsburg*, pp. 1256-1264.

Utvić, M. (2013), "The construction of reference corpus of contemporary Serbian", PhD thesis, Filološki fakultet, Univerzitet u Beogradu.

Vintar, Š. and Fišer, D. (2008), "Harvesting multi-word expressions from parallel corpora", *Proceedings of LREC'08 Conference*, *European Language Resources Association*, pp. 1091-1096.

Vitas, D. and Krstev, C. (2006), "Literature and aligned texts", in Slavcheva, M., Angelova, G. and Simov, K. (Eds), *Readings in Multilinguality*, Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia, pp. 148-155.

Vitas, D. and Krstev, C. (2012b), "Processing of corpora of serbian using electronic dictionaries", *Prace Filologiczne*, Vol. 63, pp. 279-292.

Volk, M., Graën, J. and Callegaro, E. (2014), "Innovations in parallel corpus search tools", *Proceedings of LREC'14*, *European Language Resources Association*, pp. 3172-3178.

Wu, A. and Chen, J. (2019), "Sustaining multilinguality: Case studies of two American multilingual digital libraries", *iConference 2019 Proceedings*, *iSchools*.

Xu, Y., Chen, L., Wei, J., Ananiadou, S., Fan, Y., Qian, Y., Eric, I., Chang, C. and Tsujii, J. (2015), "Bilingual term alignment from comparable corpora in English discharge summary and chinese discharge summary", *BMC Bioinformatics*, Vol. 16 No. 1, Article No. 149.

Zhang, C. and Wu, D. (2012), "Bilingual terminology extraction using multi-level termhood", *The Electronic Library*, Vol. 30 No. 2, pp. 295-308.

Further reading

Collins English Dictionary (2019), available at: www.collinsdictionary.com/dictionary/english/token (accessed 5 February 2019).

Erjavec, T., Krstev, C., Petkevič, V., Simov, K., Tadić, M. and Vitas, D. (2003), "The MULTEXT-East morphosyntactic specifications for slavic languages", in Erjavec, T. and Vitas, D. (Eds), *Proceedings of the Workshop on Morphological Processing of Slavic Languages: 10th Conference of the European Chapter (EACL'03)*, pp. 25-32.

Gambette, P. and Véronis, J. (2010), "Visualising a text with a tree cloud", *IFCS'09*, Springer, Berlin, Heidelberg, pp. 561-569.

Krstev, C. and Vitas, D. (2011), "An aligned English-Serbian corpus", in Tomović, N. and Vujić, J. (Eds), *ELLSIIR Proceedings Volume I, Faculty of Philology*, University of Belgrade, Belgrade, pp. 495-508.

Sketch Engine (2019), "Token", available at: www.sketchengine.eu/my_keywords/token/ (accessed 5 February 2019).

Vitas, D. and Krstev, C. (2012a), "Construction and exploitation of X-Serbian bitexts", in Vertan, C. and Walther v., H. (Eds), *Multilingual Processing in Eastern and Southern EU Languages: Low-Resourced Technologies and Translation*, Cambridge Scholars Publishing, Cambridge, pp. 207-227.

Corresponding author

Jelena Andonovski can be contacted at: jelenaandonovski87@gmail.com